**Diabetes Dataset**

این پروژه به چهار بخش تقسیم می­شود: 1- مدیریت داده­های گم شده. 2- حذف کردن داده­های پرت. 3- انتخاب ویژگی­ها. 4- ارائه و آموزش مدل

در ادامه، به شرح کارهای انجام گرفته در هر کدام از این سه بخش می­پردازیم.

* **مدیریت داده­های گم شده (missing values)**

در این قسمت بر اساس شرایط هر ویژگی برای داده­های گم شده آن تصمیم­گیری می­شود.

برای ویژگی “race”، پر تکرارترین مقدار را جایگزین داده­­های گم شده می­کنیم.  
برای ویژگی “gender” ، به دلیل تعداد کم داده­های گم شده، ردیف­هایی که این ویژگی آن­های گم شده است را حذف می­کنیم.

برای ویژگی “weight”، ابتدا مقدار میانه بازه­ی گفته شده را جایگزین آن بازه می­کنیم. (به عنوان مثال، به جای تمام مقادیر [0-10) عدد 5 را قرار می­دهیم.) سپس مقدار میانگین سنین گفته شده را جایگزین داده­های گم شده می­نماییم.

برای ویژگی­های diag\_1 ، diag\_2 و diag\_3 ، به دلیل تعداد کم داده­های گم شده، ردیف­های متناظر با آن­ها را حذف می­نماییم.

برای بقیه ویژگی­هایی که داده*­*­های گم شده دارند، به دلیل تعداد بالای داده­های گم شده و یا کم اهمیت بودن ویژگی­ها در پیش­بینی نهایی، آن­ها را به همان صورت باقی می­گذاریم.

* **حذف کردن داده­های پرت (outlier elimination)**

در این بخش، ردیف­هایی که احتمال وقوع مقدار یکی از ویژگی­های آن­ها از عددی معین (outlier threshold) کم­تر باشد را حذف می­کنیم.

* **انتخاب ویژگی­ها (feature selection):**

در این مرحله بر اساس مقدار Mutual Information بین ویژگی مورد نظر و ویژگی هدف، ویژگی­هایی را انتخاب می­کنیم. ابتدا مقدار Mutual Information تمام ویژگی­ها را با برچسب داده­ها (readmitted) حساب کرده و آن­ها را از بزرگ به کوچک مرتب می­کنیم. هر ویژگی­ای که MI بیش­تری با برچسب داده­ها داشته باشد، مناسب­تر است. بنابراین از بین ویژگی­ها، چند ویژگی­ای که مقدار MI بیش­تری دارند را انتخاب می­کنیم. ویژگی­های نهایی انتخاب شده در فایل preprosecced\_data.csv قابل مشاهده هستند. این فایل بعد از انجام تمام مراحل پیش پردازش داده­ها که در بخش­های قبلی به آن­ها اشاره شد بدست آمده است.

* **ارائه و آموزش مدل**

ابتدا برای استخراج ویژگی­های آموزنده­تر از داده­های عددی، یک شبکه عصبی سه لایه­ای تعریف کرده و ویژگی­های عددی را به آن ورودی می­دهیم. این شبکه عصبی دارای 7 ورودی، دو لایه مخفی با 21 گره و لایه خروجی با سه گره است. این شبکه عصبی، داده­ها را به سه گروه تقسیم­بندی می­کند. سپس بر روی هر کدام از آن سه گروه یک درخت تصمیم آموزش می­دهیم. درخت تصمیم باینری خواهد بود و بر اساس معیار GINI انتخاب می­کنیم که کدام ویژگی را گسترش دهیم. بنابر آزمایشات انجام شده، دقت این مدل برابر 75 درصد است.